Türkçe Haber Bültenleri için Dil Modelleme Yaklaşımlarının Karşılaştırması
Comparison of Language Modeling Approaches for Turkish Broadcast News

Tuncay Aksungurlu, Sıddıka Parlak, Haşim Sak, Murat Saraçlar

Elektrik-Elektronik Mühendisliği Bölümü
Boğaziçi Üniversitesi, 34342, Bebek, İstanbul, Türkiye
{tuncay.aksungurlu, siddika.parlak, murat.saraclar}@boun.edu.tr

Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi, 34342, Bebek, İstanbul, Türkiye
{hasim.sak}@boun.edu.tr

Özetçe

Bu çalışmada, çeşitli dil modelleme yaklaşımlarının, Türkçe haber bültenleri için geliştirilen konuşma tanıma sistemi üzerindeki başarımı incelenmiştir. Türkçe'nin bitişken yapısından dolayı dağarcık dışı (DD) kelime oranı yüksektir ve bu da kelime hata oranını (KHO) artırmaktadır. DD kelime sorununa çözüm olarak, kelime altı dil modelleriyle ve daha geniş dağarcıklı dil modelleriyle çalışılmıştır. Modeller istatistiksel olarak oluşturulduğundan, eğitim verisi arttıkça da başarımın artması beklenmektedir. Bu ilişkiyi incelemek amacıyla, kelime ve kelime-altı modelleri farklı veri miktarları ile eğitilmiş ve konuşma tanıma başarımları karşılaştırılmıştır.

Abstract

In this paper, we investigate the performance of several language modeling approaches on a speech recognition system for Turkish Broadcast News. The agglutinative structure of Turkish introduces a high out-of-vocabulary rate and hence increases word error rate. To eliminate out-of-vocabulary problem, we utilize various sub-word models. In addition, we experiment with high vocabulary sizes. Since the models are statistical, we expect an improvement in performance as the amount of training data increases. We build word and sub-word language models using various amounts of corpora and compare their recognition performance.

1. Giriş

Konuşma tanıma sistemlerinin en temel iki parçası akustik model ve dil modelidir. Dil modeli eğitilirken, tanıma yapılacak dilin yapısı dikkate alınır. Bir grup dilde iyi sonuç veren modeller, başka diller için uygun olmayabilir. Türkçe için geliştirilen konuşma tanıma sistemlerinde de Türkçe'nin yapısal özellikleri gözetilmelidir.

Türkçe sondan eklemeli bir dildir. Eklerin arka arkaya kullanılmasıyla aynı kökten birçok yeni sözcük elde etmek mümkündür. Bir isim ya da fiil kökünden iki milyon farklı kelime elde edilebileceği görülmüştür [1]. Örneğin Türkçe'de gövde halindeki bir kelime, bükümlü bir dil olan İngilizce'de bir grup kelimeye denk gelebilir. Bu nedenle Türkçe, kelime sayısı açısından zengin bir dildir [2] .

Konuşma tanıma için dil modeli oluşturulurken öncelikle bir dağarcık belirlenir. Sistemin DD kelimeleri tanıması olanaksızdır. Türkçe'de çok sayıda kelime olması, DD kelime oranını yükseltir ve konuşma tanıma başarımını düşürür. Türkçe'nin yapısından kaynaklanan zorlukların yanısıra, başarılı bir dil modeli için oldukça büyük bir derlemle eğitim yapılmalıdır. Sonuç olarak, üzerinde durulması gereken iki önemli nokta, DD kelimeler ve dil modelinin güvenilirliğidir.

DD kelime oranını azaltmanın bir yolu dağarcık büyüklüğünü artırmaktır. Türkçe gibi bitişken yapılı dillerde tanıma başarımının diğer dillerdekine ulaşabilmesi için çok büyük dağarcıklarla çalışılması gereklidir. Ancak geniş dağarcıkla çalışmak fazla hesap gücü ve bellek gerektirir. Dolayısıyla dağarcık büyüklüğü ancak belli bir noktaya kadar artırılabilir. Daha önce Türkçe haber programları için yapılan bazı çalışmalarda 50 bin kelimelik dağarcıklar kullanılmıştır [2] [3]. Bu çalışmada 100 bin ve 200 bin kelimelik dağarcıklar da sınanmıştır.

DD kelime sorununun bir diğer çözümü de kelime-altı birimlerle dil modelleri oluşturmaktır. Kelimeleri parçalama yöntemlerinden biri biçimbilimsel ayrıştırma ile kök ve eklerine ayırmadır. [4] 'te bu şekilde elde edilen gövde ve eklerden dil modeli oluşturulmasının yanısıra sadece gövdeler kullanılarak da modelleme yapılmıştır. [2] 'de yine biçimbilimsel ayrıştırma ile elde edilen kök ve kök sonrası (KKS) birimler kullanılmıştır. Bir başka kelime-altı birim de istatistiksel çözümleme ile elde edilen morf adı verilen birimlerdir [7]. Bu çalışmada morflarla ve KKS birimlerin başarımı sınanmıştır.

Dil modeli istatistiksel olarak oluşturulduğundan, güvenilir olması veri miktarıyla doğrudan ilişkilidir. Veri miktarının tanıma başarımı üzerindeki etkisini görmek amacıyla farklı büyüklükteki derlemlerle (25 milyon, 50 milyon, 100 milyon, 200 milyon kelime) modeller oluşturulmuştur.

Bu bildirinin ikinci kısmında, şu ana kadar toplanmış

akustik veri ve metin verisi anlatılmaktadır. Üçüncü kısımda, önce konuşma tanıma için akustik model ve dil modelinin nasıl kullanıldığından bahsedilmektedir. Daha sonra akustik modele değinilmiş ve dil modellerinden ayrıntılı olarak bahsedilmiştir. Dördüncü ve son kısımda, deney düzeneği anlatılmakta ve deney sonuçları verilmektedir.

2. Veri

2.1. Akustik Model Verisi

Akustik model eğitimi için konuşma verisine ihtiyaç vardır. Radyo ve televizyon kanallarından düzenli olarak toplanan haber kayıtları elle bölütlenip yazılandırılmaktadır. Ayrıca yazılandırmaların bir kısmı elle denetlenmiştir.

Bundan önceki çalışmalarımızda 68.6 saatlik konuşma içeren akustik veri, eklenen yeni kayıtlarla birlikte yaklaşık 200 saate ulaşmıştır. 2007 yılı Nisan ayına kadar olan programlar toplam 183.8 saat olup, eğitim verisi olarak ayrılmıştır. Sınama verisi ise 2007 yılı Nisan ayından seçilmiştir ve 3.1 saattir. Dolayısıyla aynı güne ait test ve eğitim verisi yoktur. Ayrıca, Tablo 1 'de de görüleceği gibi sınama verisi İşitme Engelli (İE) haber kayıtlarını içermemektedir. İE haber bültenleri tamamıyla temiz konuşma içerdiğinden ortalama kelime hata oranını düşürmekteydi [3]. Dolayısıyla bu çalışmadaki sınama verisi, önceki çalışmalardan daha zorlayıcıdır.

Table 1: Eğitim ve sınama verisinin kanal ve akustik ortam bilgisine göre süre analizi

Eğitim Verisi							
Kanal	f0	f1	f2	f3	f4	fx	Toplam
CNN	15.0	7.8	1.9	6.6	22.9	1.0	55.2
NTV	15.5	3.8	2.2	6.8	36.9	1.7	66.9
TRT2	5.1	1.6	0.2	2.7	8.0	0.2	17.8
İE	11.9	0	0	0	0	0	11.9
TRT1	1.2	1.4	0	0.4	2.6	0.1	5.7
VoA	17.0	0.9	4.0	3.0	1.3	0.1	26.3
Toplam	65.7	15.5	8.3	19.5	71.7	3.1	183.8
Sınama Verisi							
CNN	0.13	0.02	0	0.12	0.45	0	0.72
NTV	0.18	0.09	0.03	0.010	0.28	0.01	0.69
TRT2	0.18	0.03	0	0.22	0.46	0	0.89
VoA	0.62	0	0.04	0.08	0.07	0	0.81
Toplam	1.11	0.14	0.07	0.52	1.26	0.01	3.11

Tablo 1 'te eğitim ve sınama verisinin kanal ve akustik ortama bağlı süre analizleri verilmiştir. Akustik ortam sınıflandırmasında; f0 temiz konuşmayı, f1 spontane konuşmayı, f2 telefon konuşmasını, f3 arka planda müzik içeren konuşmayı, f4 kötü akustik koşulları, fx ise diğer konuşmaları temsil eder.

2.2. Dil Modeli Verisi

Dil modeli eğitimi için metin verisine ihtiyaç vardır. Bu araştırmada iki farklı derlem kullanılmıştır. Birincisi, akustik veri kısmında bahsedilen konuşma kayıtlarının yazılandırılmasıyla elde edilir. Bu derlem yaklaşık 1.2 milyon kelime içermektedir ve Haber Bülteni Verisi (HBV) olarak adlandırılır.

Gazete Haberi Derlemi (GHD) adı verilen diğer derlem ise, üç ana internet haber portalından geliştirdiğimiz WEB gezici yazılım ile toplanmıştır ve yaklaşık 184 milyon kelimeden oluşmaktadır. Bu derlemin temizlenmesi için biçimbilimsel çözümleyici ve bazı buluşsal yöntemler kullanılmıştır. Gazete haber derlemi 2.2 milyon farklı kelime tipi içermektedir. Biçimbilimsel çözümleyici, derlemdeki kelimelerin yaklaşık %96.7'si, farklı kelime tiplerinin ise %52.2'si için analiz sonucu dönebilmektedir. Bu sonuç derlemde çok fazla özel isim ve yazım hatası olan kelime tipi olduğunu göstermektedir. Fakat bu tip kelimelerin sıklıkları düşük olduğu için derlemde toplam sayıları nispeten azdır. Bu büyüklükteki bir derlemin dahi Türkçe'nin zengin biçimbirime sahip olması nedeniyle bütün kelime biçimlerini içermeye yetmediği gözlenmiştir.

3. Konuşma Tanıma Sistemi

Konuşma tanımada temel problem belli bir akustik sinyal verildiğinde ona karşılık gelen en olası kelime dizisini bulmaktır. Akustik sinyal öznitelik (MFCC-mel frekans kepstrum katsayısı) vektör dizileriyle temsil edilir. Her sesçik, kendinden önceki ve sonraki harfin bağlamında, Saklı Markov Modelleri (SMM) ile modellenir. Telaffuz modeli, kelimeler ve sesçikler arasında bir eşlemleme görevi görür. En sondaki dil modeli ise, o dile ait kelime dizilerini temsil eder. Akustik model, telaffuz modeli ve dil modeli ağırlıklandırılmış sonlu durum makineleri (ASDM) ile temsil edilir ve konuşma tanıma ağı bu üç modelin bileşkelenmesi (ASDM bileşke operasyonu) ile elde edilir. Bu bildiride telaffuz modeli ve dil modelinin bileşkesinden dil modeli olarak bahsedilmiştir.

3.1. Akustik Model

Akustik model, ayrıntıları Tablo 1'de verilen 183.8 saatlik konuşma verisi ile eğitilmiştir. Bütün deneylerde aynı akustik model kullanılmıştır. Öncelikle HTK yazılımı yardımıyla, öznitelik (MFCC, enerji ve türevleri) vektörleri çıkarılır. Daha sonra bu vektörler kullanılarak SMM eğitilir. SMM'de üçlü ses modelleri için karar ağaçlarıyla kümelenmiş her durum, 11 karışımlı Gauss Karışım Modelleri (GKM) ile temsil edilmektedir [3].

3.2. Dil Modeli

Kelime ve kelime altı birimlerden farklı dil modelleri oluşturulmuştur. Kelime altı birim olarak morflar ve KKS birimler kullanılmıştır. Aşağıdaki örnekte aynı söz dizisinin bu üç farklı yaklaşımla birimlerine ayrılmış hali gösterilmektedir.

```
Kelime: kesildiği andan itibaren
Morf:   kesil diği # a ndan # itibaren
KKS:    kes ildiği # an dan # itibaren
```

Dil modelinin oluşturulmasında iki veri tabanı kullanılır. GHD'den elde edilen genel dil modeli ve HBV'den elde edilen haber dil modeli doğrusal olarak aradeğerlenir ve asıl dil modeli oluşur. Aradeğerleme aşağıdaki denkleme göre yapılır.

$$P(w_3|w_2,w_1) = \lambda P_G(w_3|w_2,w_1) + (1-\lambda)P_H(w_3|w_2,w_1)$$
$$0 \leq \lambda \leq 1 \quad (1)$$

Denklemde $P_G(w_3|w_2, w_1)$ GHD ile oluşturulan dil modelini, $P_H(w_3|w_2, w_1)$ ise HBV ile oluşturulan dil modelini göstermektedir. Ölçekleme sabiti daha önceden deneysel olarak hesaplanmış ve 0.5 olarak alınmıştır [3] .

GHD'den elde edilen genel dil modelinin büyüklüğü karmaşıklık tabanlı budama ile kontrol edilir. Genel modelin örü çıktısı budama yapılmamış bir dil modeli ile tekrar değerlendirilmektedir.

Ayrıca, GHD miktarının tanıma başarımına etkisini görmek amacıyla, 200 milyonluk GHD'den rastgele 25 milyon, 50 milyon ve 100 milyonluk kısımlar seçilmiştir. Yeni veri tabanlarından elde edile genel modeller yine haber dil modeli ile aradeğerlenir ve asıl dil modeline ulaşılır.

3.2.1. Kelime Dil Modeli

Tanıma birimi olarak kelimeleri kullanmak, en temel dil modelleme yaklaşımıdır. Metin veri tabanında (GHD+HBV) en sık geçen 50 bin, 100 bin ve 200 bin kelime ile dağarcıklar oluşturulup, üçlü dil modelleri elde edilmektedir. Dağarcığın büyümesi bellek gereksinimini artırdığından, 200 bin kelimeden daha büyük dağarcıklarla çalışılmamıştır. Her bir dağarcığa ait DD kelime oranı Tablo 2 'de verilmiştir. 200 bin kelimelik dağarcığın, oldukça yüksek bir kapsam oranına (%98) sahip olduğu görülmektedir.

3.2.2. Kök ve Kök Sonrası Dil Modeli

Kök ve kök sonrası dil modelini oluşturmak için biçimbilimsel çözümleyici bir sistem kullanılmıştır. Kullandığımız çözümleyici iki aşamalı biçimbilime dayalıdır. Örnek olarak, çözümleyicinin `çocukları` kelimesi için çıktısı aşağıda verilmiştir.

```
çocuk[Noun]+lAr[A3pl]+SH[P3pl]+[Nom] (onların çocukları)
çocuk[Noun]+lAr[A3pl]+SH[P3sg]+[Nom] (onun çocukları)
çocuk[Noun]+lAr[A3pl]+[Pnon]+YH[Acc] (çocukları [gördüm])
çocuk[Noun]+[A3sg]+lArH[P3pl]+[Nom] (onların çocuğu)
```

Biçimbilimsel çözümleyici Türkçe'nin karmaşık biçimbilime sahip olması nedeniyle birden fazla sonuç döndürebilmektedir. Doğru çözümlemenin seçimi için biçimbilimsel muğlaklık giderici bir sistem kullanılması gerekmektedir, fakat bu çalışma için en az sayıda morfem içeren çözümleme seçilmiştir. Seçilen çözümleme kelimenin kök ve kök sonrasına eşleştirme yapılarak ayrılması için kullanılmıştır. Örnek olarak `çocukları` kelimesi kök ve kök sonrası model için "`çocuk +ları`" şeklinde ayrılmıştır.

Kök ve kök sonrası dil modeli için dağarcık boyutu 50,000 (41363 kök ve 8637 kök sonrası) birimle sınırlandırılmıştır.

3.2.3. Morf Dil Modeli

Kelimelerin otomatik olarak morfemlerine ayrılması işlemi, yeterince büyük bir derlem kullanıldığında, söz konusu derlemin, üzerinde çalışılan dilin yapısıyla ilgili istatistiki verileri içereceği fikrini temel alan ve eğitmensiz şekilde kelimeleri morfemlere ayırmaya çalışan algoritmalar kullanılarak da gerçekleştirilebilmektedir. Bu şekilde elde edilen morfem benzeri yapılar, bu yapıların tam anlamıyla morfem tanımına uyamayacağı göz önünde bulundurularak, morf olarak adlandırılmıştır. Yapılan çalışmada kullanılan algoritma en büyük sonsal olasılık prensibini temel alarak, derlemi meydana getiren morflardan ibaret en iyi sözlüğü elde etmeye çalışmaktadır [10]. Bu esnada "(kök+ ek*)+" kurallı ifadesiyle gösterilen Türkçe'ye özgü temel eklenim kuralı da göz önünde bulundurulmaktadır.

Yapılan çalışmada, öncelikle derlem içinde 5'ten fazla geçen kelimeler kullanılarak morflardan oluşan sözlük elde edilmektedir. Ardından tüm derlem viterbi algoritması ile, elde edilen sözlükteki morflara ayrılmaktadır. (Bu işlem esnasında uygun bölmenin sağlanabilmesi için, kaynaştırma harfleri de göz önüne alınarak, morfların yanında harflere de ufak bir olasılık atanmaktadır.) Bu şekilde elde edilen morflara ayrılmış derlemde, 142038'i kök, 42959'u ek olmak üzere toplam 184997 farklı morf elde edilmiş ve kelime başına düşen ortalama morf sayısı yaklaşık 2.71 olarak gözlenmiştir.

Ardından, diğer dil modelleriyle karşılaştırabilmek amacıyla, derlemde kullanılan kelime dağarcığı sınırlandırılmakta ve konuşma tanıma işlemine geçilmektedir. Dil modelleri oluşturulurken kullanılacak, morflara ayrılmış derlem meydana getirilirken iki farklı yaklaşım izlenmiştir. İlk olarak, derlemdeki her bir kelimenin yerine, kelimenin morflara ayrılmış hali getirilirken (morf); diğer bir yaklaşım olarak, ekler birleştirilmiş, daha önceki bölümlerde de bahsedilen "kök ve kök sonrası" yapısı kullanılmıştır (morf_KKS).

50 binlik, 100 binlik ve 200 binlik dağarcık sınırlandırması sonucu, oluşturulan morf_KKS dil modellerinde elde edilen dağarcık dışı kelime sayıları Tablo 2'de verilmiştir.

Table 2: Dağarcık Dışı Kelime Oranları (%)

Birim	50k	100k	200k
Kelime	7.3	3.9	1.9
Morf_KKS	0.4	0.1	~ 0
KKS	2.1	1.5	1.3

4. Deneyler

Bu bölümde, bahsedilen dil modeli yaklaşımları sınanmıştır. Eğitim verisi miktarının, dağarcık boyutunun ve kelime-altı birim kullanımının konuşma tanıma başarımına olan etkisi Kelime Hata Oranları (KHO) cinsinden karşılaştırılmıştır.

4.1. Deney Düzeneği

Konuşma tanıma sistemindeki dağarcık; kelimeler, morflar ve KKS birimlerden oluşur. Eğitim verisinin ve dağarcık büyüklüğünün etkisini görmek amacıyla yapılan deneylerde birim olarak sadece kelimeler kullanılmıştır. 50 bin, 100 bin ve 200 bin kelimelik dağarcıklardan ve 25, 50, 100, 200 milyonluk genel veri tabanlarından dil modelleri oluşturulmuştur.

Tablo 2 'de, kelime altı modellerin DD birim oranlarının neredeyse sabit olduğu görülmektedir. Bu sebeple, farklı birimlerle yapılan deneylerde kelime altı dağarcık boyutu 50 bin birim, kelime dağarcık boyutu 200 bin kelime olarak alınmıştır. Kelimelerde üçlü, morflarda beşli, KKS'nda ise dörtlü dil modelleri oluşturulmuştur. Dil modelleri SRILM [8] yazılımı ile oluşturulur.

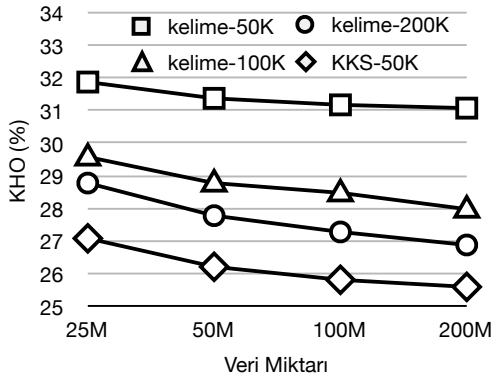

Figure 1: 50 bin, 100 bin, 200 bin kelimelik dağarcıklar için genel veri tabanı miktarına göre kelime hata oranları

Bütün deneyler için 3.1 kısmında anlatılan akustik model kullanılmıştır. Tanıma AT&T [9] kod çözücüsü ile yapılmaktadır.

4.2. Deney Sonuçları

Dağarcık boyutu ve genel veritabanı miktarı değiştirilerek elde edilen kelime tabanlı modellerin KHO oranları şekil 1'de görülmektedir. En yüksek başarım 200 bin kelimelik dağarcık için elde edilmiştir. Elde edilen en düşük KHO % 26.9'dur. Beklendiği gibi, genel veri tabanı miktarı arttıkça da başarım artmaktadır. Veri miktarı 25 milyondan 200 milyona yükseltilirken en fazla iyileşmenin 200 bin kelimelik dağarcıkta olduğu görülmüştür (%2). Buna karşın 50 bin kelimelik dağarcığın başarımının aynı oranda iyileşmesi için çok daha fazla veri gerekmektedir.

Table 3: Farklı birimlere ait kelime hata oranları (%)

Birim	KHO
Kelime	26.9
Morf	25.9
Morf_KKS	25.3
KKS	25.6

Yine aynı grafikte, dağarcığı 50 bin birim olan KKS modelinin bütün kelime modellerinden daha iyi sonuç verdiği görülmektedir. Eh düşük KHO % 25.6 olarak hesaplanmıştır. 200 milyonluk derlem üzerinde, 50 bin birimlik dağarcık ile, morf_KKS tabanlı dil modeli kullanılarak yapılan konuşma tanıma deneyinde ise KHO % 25.8 olarak elde edilmiştir. Her iki kelime altı model de kelime modelinden daha iyi sonuçlar verirken, başarımları biribirine yakındır.

Tablo 3 'te, 200 milyonluk derlem üzerinde farklı birimler için en iyi kelime hata oranları görülmektedir. En iyi sonuçlar, bölütlemeden bağımsız olarak kök ve kök sonrası modelleme ile elde edilmiştir. Dağarcık genişliğinin küçük olduğu durumlarda kelime altı modellerle çalışmak oldukça büyük kazançlar sağlamaktadır. Ancak, dağarcık boyutu artırıldığında kelime modelleri de karşılaştırılabilir bir başarım sergilemektedir.

5. Teşekkür

Yazarlar metin veritabanı için Sabancı Üniversitesi'ne ve ODTÜ'ye, konuşma tanıma düzeneği ve akustik model için Ebru Arısoy'a ve Doğan Can'a, ASDM yazılımları için AT&T - Labs Research'e teşekkür ederler. Bu çalışma 105E102 ve 107E261 nolu TÜBİTAK projeleri ile 05HA202 ve 07HA201D nolu Boğaziçi Üniversitesi BAP projeleri kapsamındadır. Birinci ve üçüncü yazar TÜBİTAK BİDEB 2210 ve 2211 programları içerğinde desteklenmektedir.

6. Kaynakça

[1] Güngör T., Sak H., Saraçlar M., "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus", Language Resources and Evaluation, 2008.

[2] Arısoy E., Sak H., Saraçlar M., "Language Modeling for Automatic Turkish Broadcast News Transcription", International Conferance on Spoken Language Processing - Interspeech, 2007.

[3] Arısoy E., Saraçlar M., "Türkçe Haber Programları için Konuşma Tanıma", Sinyal İşeleme ve Uygulamaları Kurultayı, 2007.

[4] Çiloğlu T., Çömez M., Şahin S., "Takılı Bir Dil Olarak Türkçe İçin Dil Modelleme", Sinyal İşeleme ve Uygulamaları Kurultayı, 2004.

[5] Young S., Ollason D., Valtchev V., Woodland P., "The HTK book (for HTK version 3.2.)", March 2002. http://htk.eng.cam.ac.uk/

[6] Creutz, M. and Lagus, K., "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor", Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005. http://www.cis.hut.fi/projects/morpho/

[7] Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimaki, T., Pylkkonen, J., Alumae, T., and Saraclar, M., 2006, "Unlimited vocabulary speech recognition for agglutinative languages", In Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2006. New York, USA, June 5-7, 2006.

[8] Stolcke, A., "SRILM - An extensible language modeling toolkit", Proceedings of the International Conference on Spoken Language Processing, 2002, 901-904. http://www.speech.sri.com/projects/srilm/

[9] Mohri M. and Riley, M. D., "DCD Library - Speech Recognition Decoder Library". AT&T Labs - Research. http://www.research.att.com/sw/tools/dcd/

[10] Creutz M. and Lagus K., "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text", In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR05), pp. 106–113, Espoo, Finland, June 2005.